SEGMENTATION OF INDUS TEXTS[1]

Nisha Yadav[@], M N Vahia[@], Iravatham Mahadevan[#], Hrishikesh Joglekar*

[@] *Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005*
[#] *Indus Research Centre, Roja Muthiah Research Library, Chennai*
**Khagol Mandal, Mumbai*

Abstract: *We adopt a comprehensive approach to segment the Indus texts using statistically significant signs and their combinations in addition to all the texts of length 2, 3 and 4 signs. We find that we can segment 88% of Indus texts (of length 5 and above) by this method and hence it can be suggested that the texts of 5 or more signs can actually be seen as permutations of other frequent sign-combinations or smaller texts (of length 2, 3 or 4 signs). The results of the segmentation process are in agreement with our earlier results (Yadav et. al, 2008, henceforth referred to as Paper 1) where we show the importance of 2, 3 and 4 sign combinations as important units of information. We do not assume anything regarding the content of the script and the work is purely based on the structural analysis of Indus Texts.*

1.0 Dataset: We use electronic concordance of Mahadevan (1977), henceforth referred to as M77 (For details see Paper 1). M77 records 417 unique signs[2] in 3573 lines of 2906 texts. We remove texts that can have potentially ambiguous reading. We create an Extended Basic Unique Data Set (EBUDS) by removing all texts containing lost, damaged or illegible passages marked by diagonal lines and doubtfully read signs marked by asterisk. All texts from multi-lined sides are also removed. However, we assume that in objects where writing is found on several sides, the text on each side is independent of text on

[1] Address for correspondence: y_nisha@tifr.res.in

[2] The serial number of the signs used in this paper is as given by Mahadevan in his concordance (1977). As a convention followed in the present paper, the texts depicted by pictures are to be read from right to left, whereas the texts represented by just strings of sign numbers are to be read from left to right.

other side(s). We retain texts from those sides of multisided objects which have only one line of text. Texts appearing more than once are taken only once. We do not take into account the variation due to archaeological context of sites, stratigraphy and the type of objects on which the texts are inscribed.

The unit of textual analysis for the study of distributional statistics is a line of text. There are two reasons why it is not possible to consider the whole text on a single side as a unit for this purpose. Firstly, there is no way of knowing beforehand whether different lines of an inscription appearing on the same object or even on the same side have continuity of sequence or to be regarded as separate texts. Secondly, it is not possible to ascertain beforehand the real order (if any) of the lines of text appearing on the same side (Mahadevan, 1977, p. 10).

EBUDS contains 1548 texts. In EBUDS, 40 signs out of 417 present in the Sign List of Mahadevan do not make their appearance. Out of these removed 40 signs, one sign (sign number: 374) appears 9 times, one sign (sign number: 237) appears 8 times, two signs (sign numbers: 282, 390) appear 3 times, three signs (sign numbers: 324, 376, 378) appear twice and thirty-three signs appear only once in M77. Hence all these 40 signs are rarely occurring signs and their absence in EBUDS does not significantly alter the patterns of writing.

2.0 Segmentation Approach

The Indus texts can be segmented by any of the following methods.

a) Comparing two texts[3]: Two texts which are identical except for a few signs at the beginning or end can be compared and it can help us extract the segments (Mahadevan, 1978).

[3] The term "text" implies complete line of text of Indus signs and EBUDS consists of 1548 such line of texts with variable lengths (1 to 14 signs).

b) Using frequent combinations of signs[4]: There are some frequent combinations of two-signs, three-signs etc. which can be treated as segments or identifiable units merely by their frequent rate of occurrence (Mahadevan, 1978). In Paper 1 we had shown that their frequency is far greater than would be expected by random chance.

c) Using sign-pair frequencies: The strongest and weakest junction points in a text based on the frequency of adjacent sign-pairs can be used for segmentation (Mahadevan, 1978).

d) Using Single Signs: Single signs falling in the categories of frequent beginners, frequent enders, and frequent auxiliary enders can be used to segment these texts.

All these methods are cumulative and overlapping. Hence, it becomes critical to decide which method should be given priority over others for the process of segmentation so that we end up with meaningful segments.

We adopt a step by step approach to segment the Indus texts of 3 or more signs. We have used statistically significant units (combination of signs or single signs) in addition to all texts of length 2, 3 and 4 for the process of segmentation. The following section discusses the various segmentation units in detail.

[4] "Frequent combination of signs" is a combination of Indus signs present anywhere in the text. They are characterised by their frequent rate of occurrence in distinct Indus texts. They can be viewed as part of a complete Indus text but sometimes that combination does appear as a complete Indus text. One example of such frequent sign combination is "267, 99" occurring 168 times in the complete corpus of EBUDS. It appears as an independent text once in EBUDS. Another example of such frequent sign combination is the sequence "336, 89, 211".

3.0 Segmentation Units:

Segmentation units are defined as the texts (of 2, 3 or 4 signs) and other statistically significant units used for segmentation of Indus texts. The segmentation units are

1) Two-sign, Three-sign and Four-sign Texts (Table 1)
2) Frequent sign combinations of 2, 3 and 4 signs (Tables 2-11).
3) Single Signs: Text Beginners, Text Enders and as Auxiliary Text Enders (Tables 12-14).

Each of these units is explained below in detail.

3.1 Two-sign, Three-sign and Four-sign Texts

The two-sign, three-sign and four-sign texts that appear as complete texts in EBUDS form the first set of segmentation units. Table 1 gives the number of texts of various lengths (in terms of number of signs) in EBUDS.

Table 1: Number of texts of lengths 1 to 14 in EBUDS

No. of Signs in the Text	No. of Texts (EBUDS)
1	69
2	189
3	284
4	263
5	296
6	195
7	133
8	59
9	26
10	21
11	9
12	1
13	1
14	2
Total	**1548**

As can be seen from table 1 EBUDS has 189 texts of length 2 (P1 to P189), 284 texts of length 3 (T1 to T284) and 263 texts of length four (Q1 to Q263).

3.2 Frequent Sign Combinations of 2, 3 and 4 signs (Beginner, Ender and Middle)

Frequent sign combinations of 2, 3 and 4 signs that appear predominantly (≥ 50 % of times) at beginning, ending or middle positions in Indus Texts (Tables 3-11) form the second set of segmentation units.

Table 2: Selection Criteria of 2, 3 and 4 sign combinations used as segmentation units

Sl. No.	Sign-Combination	Maximum Frequency	Total Frequency cut-off (*)
1	Two-sign	168	≥ 20
2	Three- Sign	34	≥ 10
3	Four-sign	16	≥ 4

*The cut-off for total frequency of occurrence is selected by taking into consideration frequency of occurrence of most frequently occurring combination in the respective category. The beginner, middle and ender combinations of 4, 3, and 2 signs are given in tables 3-11 respectively. These were used for the segmentation of the texts already segmented using two-sign, three-sign and four-sign texts (section 3.1).

Table 3: Beginner Four-sign Combinations

Sl. No.	Beginner Four-sign combination					Total Frequency of occurrence	Percentage of occurrence as beginner	Marker
	Sign No.				Pictorial version			
1	391	99	87	59		4	100	QB1
2	267	99	87	59		9	67	QB2

Table 4: Middle Four-sign Combinations

Sl. No.	Middle Four-sign Combination					Total Frequency of occurrence	Percentage of occurrence at middle	Marker
	Sign No.				Pictorial version			
1	67	51	130	149		4	100	QM1
2	65	72	336	89		4	100	QM2
3	171	59	336	89		6	83	QM3

Table 5: Ender Four-sign Combinations

Sl. No.	Ender Four-sign Combination					Total Frequency of occurrence	Percentage of occurrence as ender	Marker
	Sign No.				Pictorial version			
1	98	178	389	15		5	100	QE1
2	99	171	8	342		4	100	QE2
3	59	249	169	342		4	100	QE3
4	343	249	162	342		4	100	QE4
5	59	336	89	211		4	100	QE5
6	70	336	89	211		4	100	QE6
7	99	336	89	211		4	100	QE7
8	72	336	89	211		5	80	QE8
9	175	230	53	342		5	80	QE9
10	336	89	216	254		4	75	QE10
11	65	67	342	1		4	75	QE11
12	51	130	149	342		16	69	QE12
13	72	182	293	342		4	50	QE13
14	67	171	8	342		4	50	QE14

Table 6: Beginner Three-sign Combinations

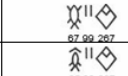
Sl. No.	Beginner Three-sign Combination				Total Frequency of occurrence	Percentage of occurrence as beginner	Marker
	Sign No.			Pictorial version			
1	293	123	343	 343 123293	25	100	TB1
2	267	99	67	 67 99 267	14	93	TB2
3	267	99	65	 65 99 267	12	92	TB3
4	267	99	87	87 99 267	14	79	TB4

Table 7: Middle Three-sign Combinations

Sl. No.	Middle Three-sign Combination				Total Frequency of occurrence	Percentage of occurrence at middle	Marker
	Sign No.			Pictorial version			
1	99	87	59	 59 87 99	16	100	TM1
2	72	336	89	 89 336 72	14	86	TM2
3	51	130	149	 149 130 51	19	79	TM3
4	53	171	59	 59 171 53	10	60	TM4

Table 8: Ender Three-sign Combinations

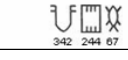
Sl. No.	Ender Three-sign Combination				Total Frequency of occurrence	Percentage of occurrence as ender	Marker
	Sign No.			Pictorial version			
1	25	245	245	 245 245 25	10	90	TE1
2	336	89	211	 211 89 336	34	88	TE2
3	249	162	342	 342 162249	24	83	TE3
4	403	87	342	 342 87 403	16	81	TE4
5	130	149	342	 342 149 130	16	75	TE5
6	171	8	342	 342 8 171	19	74	TE6
7	178	389	15	 15 389 178	11	73	TE7
8	249	169	342	 342 169249	20	70	TE8
9	67	244	342	 342 244 67	12	67	TE9

Table 9: Beginner Two-sign Combinations

Sl. No.	Beginner Two-sign Combination			Total Frequency of occurrence	Percentage of occurrence as beginner	Marker
	Sign No.		Pictorial version			
1	293	123		40	100	PB1
2	150	123		22	95	PB2
3	391	123		20	95	PB3
4	391	99		56	91	PB4
5	267	99		168	86	PB5
6	267	402		20	55	PB6

Table 10: Middle Two-sign Combinations

Sl. No.	Middle Two-sign Combination			Total Frequency of occurrence	Percentage of occurrence at middle	Marker
	Sign No.		Pictorial version			
1	99	67		26	100	PM1
2	123	343		25	100	PM2
3	99	87		24	100	PM3
4	99	65		22	100	PM4
5	99	387		22	100	PM5
6	249	169		20	90	PM6
7	336	89		75	89	PM7
8	249	162		34	85	PM8
9	171	59		36	81	PM9
10	87	59		39	79	PM10
11	171	8		21	76	PM11
12	403	87		20	75	PM12
13	65	67		27	74	PM13
14	51	130		27	70	PM14

Table 11: Ender Two-sign Combinations

Sl. No.	Ender Two-sign Combination			Total Frequency of occurrence	Percentage of occurrence as ender	Marker
	Sign No.		Pictorial Version			
1	342	176		59	97	PE1
2	89	211		34	91	PE2
3	59	211		31	90	PE3
4	342	1		48	90	PE4
5	347	342		56	89	PE5
6	53	342		24	88	PE6
7	87	342		20	85	PE7
8	162	342		25	84	PE8
9	296	342		21	81	PE9
10	169	342		22	73	PE10
11	8	342		58	72	PE11
12	245	245		33	61	PE12
13	48	342		38	53	PE13

3.3 Using Single Signs (Beginner, Ender and Middle)

Text Enders, Text Beginners and Auxiliary Text Enders form the third set of segmentation units. Based on the percentage of occurrence at the beginning, middle or end of texts, we categorise the most frequent signs as Text Enders, Text Beginners and Auxiliary Text Enders. Each of these is explained below.

i) *Text Beginners*: Text Beginners are defined as signs appearing predominantly (≥ 50 % of times) at the beginning of texts (Table 13).

ii) *Text Enders*: Text Enders are defined as signs appearing predominantly (≥ 50 % of times) at the end of texts (Table12).

iii) *Auxiliary Text Enders*: Auxiliary Text Enders are defined as signs appearing predominantly (≥ 50 % of times) at the middle of texts (Table 14), generally preceded by Text Beginners.

These are listed in tables 12-14.

Table 12: Text Ender Signs

Sl. No	Ender Sign		Total Frequency of occurrence	Percentage of occurrence as ender	Marker
	Sign No.	Pictorial version			
1	12		50	86	E1
2	176		162	84	E2
3	211		137	83	E3
4	254		51	80	E4
5	1		61	77	E5
6	15		61	75	E6
7	342		715	73	E7
8	169		106	58	E8

Table 13: Text Beginner Signs

Sl. No.	Beginner Sign		Total Frequency of occurrence	Percentage of occurrence as beginner	Marker
	Sign No.	Pictorial version			
1	267		211	78	B1
2	391		128	71	B2
3	293		90	59	B3

Table 14: Auxiliary Text Ender (Middle) Signs

Sl. No	Auxiliary Ender Sign		Total Frequency of occurrence	Percentage of occurrence at middle	Marker
	Sign No.	Pictorial version			
1	99		377	98	AE1
2	123		127	98	AE2

4.0 Method employed in segmenting Indus texts

We focus on segmenting the 734 texts of 5 or more signs to see if they are composites made of smaller information units. The steps followed in the segmentation process are explained below (Fig. 1)

STEPS FOR SEGMENTATION OF AN INDUS TEXT

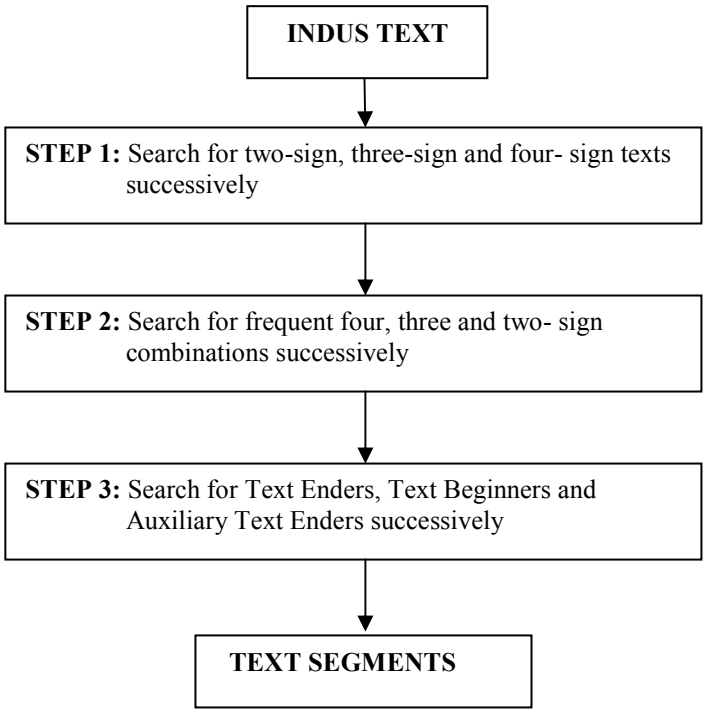


Fig. 1: Steps for segmentation of an Indus text

STEP 1: *Search for two-sign, three-sign and four-sign texts successively*

We start with 189 two-sign texts as basic segments and search the whole dataset of 1290 texts (with 3 or more signs only) for these basic segments, marking them using different markers wherever found.

This is followed by similar search for 3 and 4 sign texts respectively on the resultant dataset (dataset which had been searched for two-sign texts). We give importance to smaller texts (here two-sign texts) over three and four sign texts because a larger text could be a combination of one or more smaller units and the independent occurrence of the smaller unit increases the probability of smaller unit being a unit of information. The segmentation process is executed as follows:

- We take all stand-alone texts of length 2, 3 and 4 as complete units of information.
- For this analysis, we do not take single signs which appear solo. There are 69 signs in EBUDS that appear solo and they may artificially split grammatically significant units of information. We know that there are several cases where a given sign appears solo a few times, but appears with a *specific* other sign far more frequently indicating that its two-signed appearance carries far greater significance. Hence as an approximation, we begin with texts of length 2 or more.
- We segment larger texts using the two-sign, three-sign and four-sign texts successively.
- We split first with two-sign texts which represent *smallest bits of information*. At the end of this step 45% of texts (of length 5 and above) remain unsplit.

STEP 2: *Search for frequent four, three and two-sign combinations successively*

The resultant dataset (from step 1) is then segmented using frequent 4, 3 and 2 sign combinations successively. These are listed in tables 3-11. The segmentation process is executed as follows:

- In this step, we search for frequent sign combinations.
- Since these are *not* found stand-alone very often, they may or may not be complete. However, irrespective of whether they are completely stand-alone or not, they do represent identifiable units of information which can be islanded from its neighbourhood of signs. We therefore search for such frequent sign combinations in the resultant data set (from step 1).
- Unlike step1, we reverse the order while searching for frequent sign combinations as four, three and two successively, since a four-sign frequent combination is more likely to be a significant unit than a two-sign frequent combination.
- At the end of this step 23% of texts or segments (of length 5 and above) remain unsplit.

STEP 3: *Search for Enders, Beginners and Auxiliary Enders successively*

The Indus texts after undergoing segmentation using 2, 3 and 4 sign texts (step 1) and then by frequent sign combinations (step 2) are subjected to further segmentation using statistically significant Text Ender, Text Beginner and Auxiliary Text Ender (Middle) signs.

- In case a text or segment of 5 or more signs is not segmented by step 1 and step 2, we try segmenting the same based on frequently found text beginners or text enders.

- At the end of this step 17% of texts or segments (of length 5 and above) remain unsplit.
- We then use 'auxiliary' text enders that commonly appear just after the standard text beginners, for segmentation, and at the end of this step 12% of texts or segments (of length 5 and above) remain unsplit.

The complete procedure results in splitting 88% of the texts (of length 5 and above) in EBUDS. The results are tabulated in table 15 (Fig. 2).

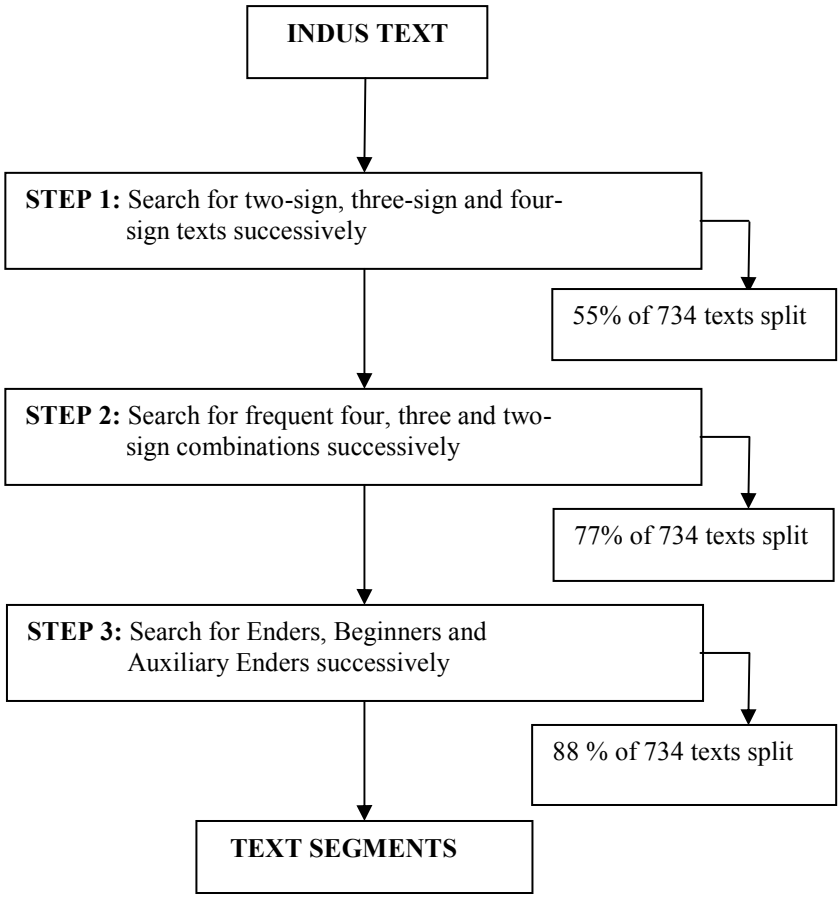


Fig.2. Results of Segmentation process

Table 15: Results of segmentation starting with 734* texts

Sl. No.	Segmentation unit	No. of segments of length 5 and above		
		No. of segments of length 5 and above remaining un-split	Split % of 734 texts taken for segmentation	Un-split % of 734 texts taken for segmentation
1	Texts of length 2, 3, 4	334	55	45
2	Frequent Combination of length 4, 3,2 (Beginners and Enders only)	250	66	34
3	Freq. combination of length 4, 3, 2 (Middle)	168	77	23
4	By Text Enders	141	81	19
5	By Text Beginners	130	83	17
6	By Auxiliary Text Enders	89	88	12

* There are 734 texts of length 5 and above in EBUDS

5.0 Results

In table 16, we list out the number of segments of various lengths after segmentation. The length vs. frequency of texts or segments is given in Fig. 4. EBUDS before and after segmentation is given in Fig. 5.

Table 16: Number of texts of Lengths 1 to 14 in EBUDS before and after segmentation

No. of Signs	No. of Texts (EBUDS)	Number of segments (EBUDS after segmentation)
1	69	630
2	189	1638
3	284	588
4	263	208
5	296	52
6	195	26
7	133	7
8	59	3
9	26	1
10	21	0
11	9	0
12	1	0
13	1	0
14	2	0
Total	1548	3153

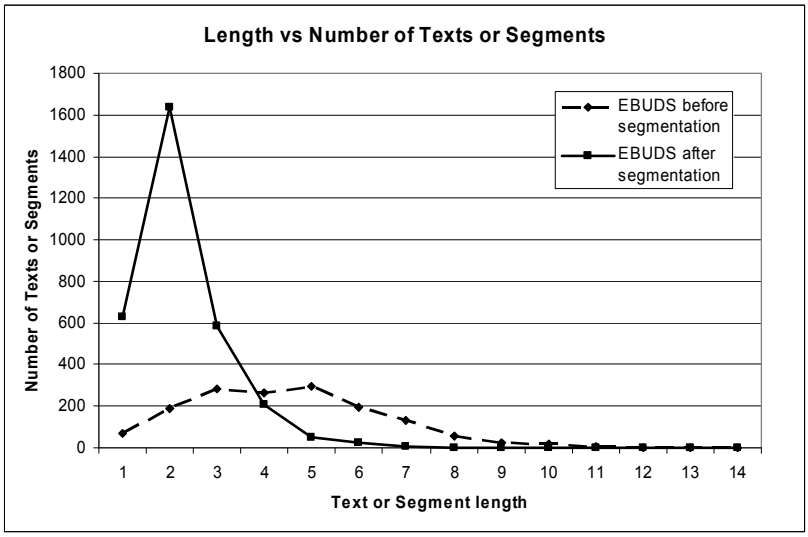


Fig. 4: Segment Length vs. Segment Frequency in EBUDS before and after segmentation

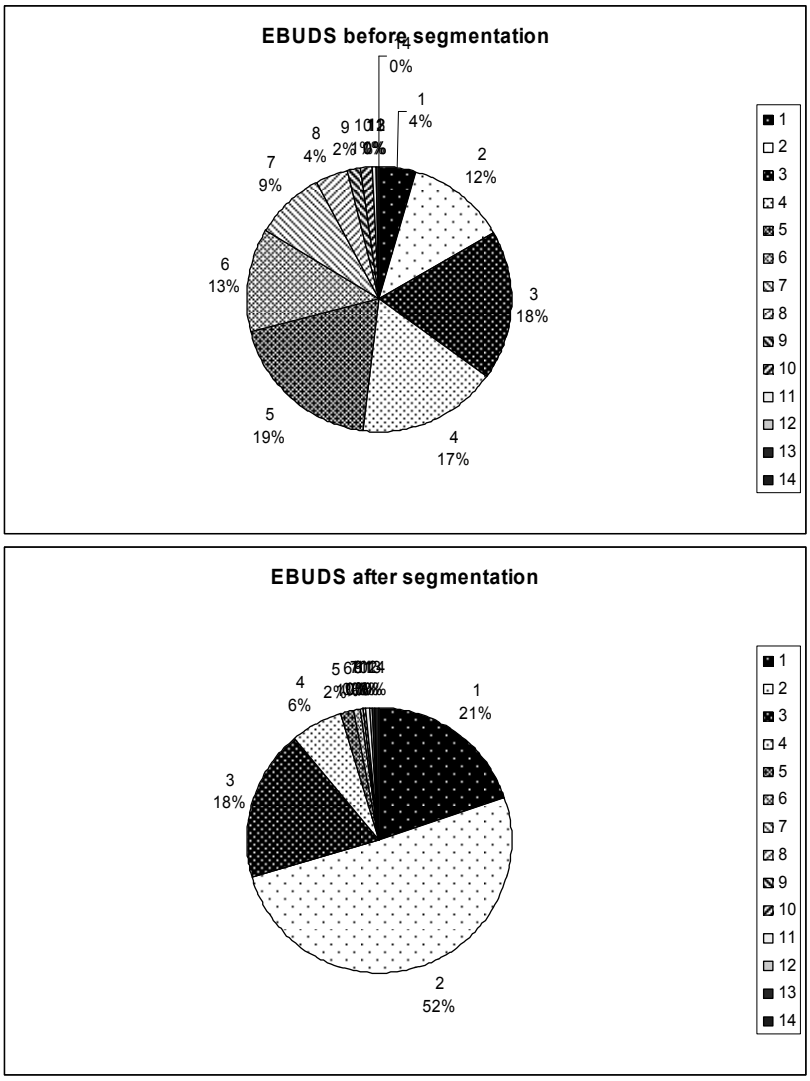


Fig. 5: EBUDS before and after segmentation

It must be noted that if these units i.e. 2, 3 and 4 sign texts do not significantly contribute to the process of segmentation of larger Indus texts then considering them as segmentation units becomes questionable. However, *finding them as part of larger Indus texts in a frequent manner justifies the nature of these 2, 3 or 4 sign texts as consciously written important pieces of*

information. Table 18 lists most frequent segments (Texts or Frequent sign combinations occurring in EBUDS after segmentation).

Table 18

Sl. No.	Frequent segments in EBUDS after segmentation					Total Frequency of occurrence
	Marker	Sign No.			Pictorial Version	
1	P86		267	99	 99 267	167
2	P169		8	342	 342 8	57
3	PB4		391	99	 99 391	46
4	P148		48	342	 342 48	37
5	T148	336	89	211	 211 89 336	34
6	PE5		347	342	 342 347	28
7	PM10		87	59	 59 87	26
8	PM7		336	89	 89 336	25
9	TB1	293	123	343	 343 123293	24
10	P42		162	342	 342 162	24
11	PB2		150	123	 123150	21
12	P43		169	342	 342 169	21
13	P100		296	342	 342 296	20

Table 19 lists few examples of Indus Texts segmented using this method. The number in the first column is the object number (in M77) of the complete text. The number at the bottom of each smaller collection of signs is the object number (in M77) on which that segment appears.

Table 19: Few Examples of Segmentation

Object No.	Segments of Text						
1232:	P148	P86					
	 342 48	 99 267					
	1279	4441					
4254:	P53	T148	P116	PM9	389		
	 14 2	 211 89 336	 342 341	 59 171	 389		
	2371		1226				
2537:	P41	PM14	67	PM9	389	344	PB1
	 342 149	 130 51	 67	 59 171	 389	 344	 123293
	8001						
2461:	T94	326	87	P131	178		
	 245 245 25	 326	 87	 15 389	 178		
	1437			2682			

6.0 Conclusion

It is possible to segment 88% of Indus Texts into segments of length 4 and below by using statistically significant signs and their combinations in addition to all the texts of length 2, 3 and 4. Based on the analysis of the segments obtained as a result of the above segmentation process we draw the following conclusions:

1) Many frequent sign combinations make their appearance as independent texts and hence considering these frequent sign combinations as units of information is justified for segmenting these texts.
2) The frequent sign-combinations which appear as independent texts are those that most often occur at the beginning or end of Indus Texts. The frequency of occurrence of a frequent sign combination which often comes at the middle of Indus text, as an independent text, is quite low.

3) The graph of segment length vs. segment frequency (Fig.4) again shows the importance of 2, 3 and 4 sign segments (Paper1) that are far more frequent than the large segments and hence larger texts can be seen as a combination of small segments of information.
4) The Indus texts after segmentation can be viewed as permutations of the identifiable units (segments) of 2, 3 or 4 signs. The identifiable units may or may not be standalone (or complete) pieces of information.

The nature of Indus writing that emerges from this and earlier work (Paper 1) is as follows. The written material is ordered in a statistically significant manner. The usage of signs is not uniform and nor is their pairing. There are clearly some important signs that appear far more often than other signs. Similarly, there are sign combinations that also appear to be intentionally paired. These aspects were discussed in Paper 1.

The study presented here indicates that these frequent sign-combinations have an additional property. These frequent sign-combinations appear to be placed within a larger text in specific sequencing. The standalone texts and most frequent signs and sign combinations are in fact parts of larger texts. This indicates that larger texts are a conglomeration of smaller texts or information units.

Acknowledgement

We wish to thank Dr. K. Samudravijaya for his enthusiastic support for this work. We wish to acknowledge the generous assistance of Jamsetji Tata Trust for this work. We also wish to acknowledge the kind hospitality of the Indus Research Centre of the Roja Muthiah Research Library for this work.

References

Mahadevan, I., 1977, *THE INDUS SCRIPT Texts, Concordance and Tables*, Memoirs of the Archaeological Survey of India No. 77.

Mahadevan, I., 1978, Recent Advances in the Study of the Indus Script, Puratattva, 9, pp 34 – 42.

Yadav, N., Vahia, M.N., Mahadevan, I., Joglekar, H., 2008, A statistical approach for pattern search in Indus writing, to appear in International Journal of Dravidian Linguistics, January 2008 (Paper 1).